

An efficient stratified-based ground motion selection for cloud analysis

Mohsen Zaker Esteghamati

Graduate Student, Dept. of Civil and Environmental Engineering, Virginia Tech, Blacksburg, USA

Qindan Huang

Associate Professor, Dept. of Civil Engineering, University of Akron, Akron, USA

ABSTRACT:

In the quantitative seismic risk assessment of structures, cloud analysis has been widely used due to its simplicity to obtain the conditional probability of structural response exceeding a certain level, conditioned on the ground motion intensity. The accuracy of this analysis relies on the selected ground motion records in terms of seismic hazard levels and the number of records. This paper presents an adaptive ground motion selection approach with a stratified sampling scheme to reduce the number of required analysis and to accurately capture structural response with a desired level of confidence. The stratified sampling scheme is used to obtain enough data points from each hazard level in an iteration fashion, while the formulation of the seismic demand model of interest is determined using a Gaussian mixed model clustering algorithm at each iteration. The proposed ground motion selection approach is applied to obtain seismic demand hazards of a non-linear single-degree-of-freedom system and the results are compared to a site-consistent model. The results show that the proposed selection method is efficient, particularly at near collapse limit states.

Cloud analysis is a popular numerical approach for evaluating seismic structural performance, where the structure is numerically modeled and subjected to a group (or cloud) of ground motion (GM) records, and time-history of engineering demand parameters (EDPs) of interest is recorded. Based on the numerical results, the EDP prediction is then typically modelled using seismic intensity measure (IM) through linear regression as follows:

$$\ln(EDP) = b_0 + b_1 \ln(IM) + \sigma \varepsilon \quad (1)$$

where b_0 and b_1 are the model parameters of the regression model fitted to the IM-EDP data in logarithm space, σ is the standard deviation of the model error, and ε is a random variable following standard normal distribution. The regression equation is then used to derive seismic fragility, which is the conditional cumulative probability of

EDP exceeding a certain level, edp , for a given IM as follows:

$$\begin{aligned} P(EDP > edp|IM) \\ = 1 - \Phi \left[\frac{edp - b_0 - b_1 \ln(IM)}{\sigma} \right] \end{aligned} \quad (2)$$

where Φ is the cumulative standard normal distribution. Lastly, the mean annual frequency (MAF) of EDP exceeding a given level, λ_{EDP} , can be obtained by integrating the convolution of seismic fragility and IM hazard curve as follows:

$$\lambda_{EDP} = \int_{IM} P(EDP > edp|IM) \cdot \left| \frac{d\lambda_{IM}}{dIM} \right| dIM \quad (3)$$

where λ_{IM} is MAF of IM exceeding a particular level of IM, also denoted as IM hazard. Although the cloud analysis concept is inherently simple and has been widely adopted, attention still needs be given to examining the regression assumptions (e.g.

residual heteroscedasticity) and its strong sensitivity to the selected GMs (Jalayer et al. 2017).

The current literature on cloud analysis has mainly focused on the model development of EDPs (e.g., Jalayer et al. 2015; Zareian et al. 2015), and there is a lack of standard guideline for the selecting GM records used in the cloud analysis. Among the few available studies, Bradley et al. (Bradley et al. 2015) have compared annual exceeding of EDP based on a stratified GM selection method to a direct hazard consistent benchmark and concluded that while two methods agree reasonably well for IMs with strong correlation to the EDP, the bin sizes used in the GM selection method and IM choice significantly affect the results. Miano et al. (2017) suggested that a wide range of IM should be considered for cloud analysis, where a “significant portion” (e.g. 30%) of records should push structure to its life safety limit state and “too many” (e.g. 10%) records should not be included from the same earthquake event; however, these conditions are subjective as they depend on analyst experience.

The goal of this paper is to provide a procedure of GM selection for the cloud analysis, which can maintain the EDP prediction accuracy with a minimum number of time-history analysis needed. In this regard, an adaptive GM selection approach with a stratified sampling scheme is developed. The proposed approach is an iteration process and stops when the EDP model is stabilized. In addition, a clustering algorithm is incorporated in the GM selection process to determine the appropriate EDP model formulation. As such the proposed approach ensures the accuracy of the EDP model and also preserves the appropriate variability at different IM levels. Lastly, this study explores the application of the adaptive GM selection in an analysis when more than one EDP models need to be developed and some recommendations are provided.

1. STRUCTURAL BENCHMARK DESCRIPTION

In order to examine the effectiveness of the proposed GM selection, a nonlinear single-degree-of-freedom (SDOF) system is considered to develop a benchmark. Figure 1 shows the

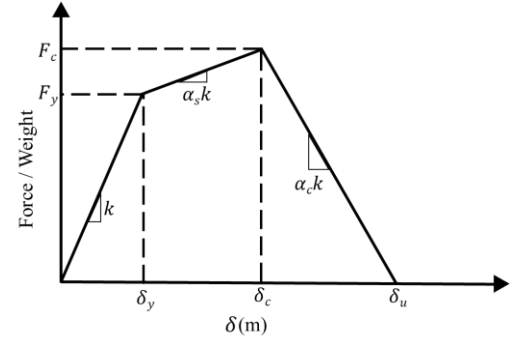


Figure 1: The backbone curve of the considered SDOF model.

backbone curve of the SDOF. The ratio of capping and ultimate displacement to the yield displacement (i.e., δ_c/δ_y and δ_u/δ_y , respectively) are taken as 2 and 8, respectively. Yield force (F_y) is assumed as 25% of weight and the critical damping is assumed to be 2%. The SDOF's period is set to be 1s. A peak-oriented hysteretic behavior is defined to account for the stiffness and strength deterioration of the SDOF system.

The “point-of-Comparison” notion is used to develop a benchmark model to assess the GM selection method, where “point-of-Comparison” refers to a regression-based EDP model that is obtained by performing a very large number of structural analysis (Watson-Lamprey 2007; Kwong 2015). In this study, the benchmark is obtained by subjecting the SDOF to 5000 site-specific simulated GMs. The simulated GMs are taken from a site with soft soil and $V_{s200}=200$ m/s in Christchurch, New Zealand with latitude and longitude of 43.53 and 172.63, respectively. More details of the simulated GMs can be found in Bradley et al. (2015).

2. SEISMIC DEMAND MODELS

While conventionally the EDP prediction is developed using linear regression over EDP and a selected IM in logarithm space, research (e.g., Bai et al. 2009) found that using bi-linear models is able to capture structural nonlinear behavior more accurately. To illustrate the importance of adopting piece-wise EDP models, with the benchmark data Figure 2 show the three different EDP models for displacement prediction (i.e., linear, bi-linear, tri-

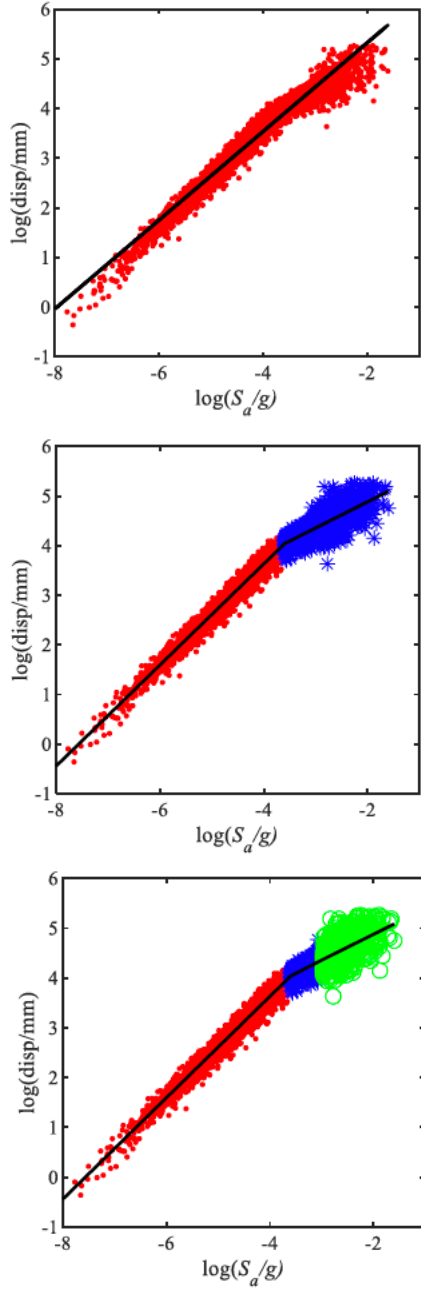


Figure 2: Comparison of EDP models: linear model (top), bi-linear model (middle), and tri-linear (bottom)

linear) using spectrum acceleration at structural fundamental period, S_a , as the predictor. The top plot in Figure 2 shows the prediction based on one linear regression line with a constant slope and variation. On the other hand, using piece-wise regression with different slopes and standard deviations of model error for each segment, the

structural behavior is better predicted, as shown in the middle and bottom plots in Figure 2. In the tri-linear models (shown in the bottom plot in Figure 2), the slopes of the last two segment are nearly the same, although the standard deviations of the model error are different.

Figure 3 shows the MAF curves of EDP based on the three models. It can be seen that the EDP results based on bi-linear and tri-linear EDP models are very similar, and they are different from the EDP curve based on the linear EDP model particular for larger EDP values. This result indicates that using piece-wise regression is necessary; however, using bi-linear EDP model is accurate enough to obtain the EDP MAFs. A similar observation has been found for peak acceleration responses.

2.1. Clustering algorithms to determine EDP model formulation

The proposed adaptive GM selection approach is an iteration process: the GM records are added in a small number at a time for nonlinear time-history analysis. The criteria for stopping selecting additional GM records is based on the current EDP model developed using all the GM records selected in the current stage. If the accuracy of EDP model is obtained, the iteration will then stop to avoid unnecessary nonlinear time-history analysis.

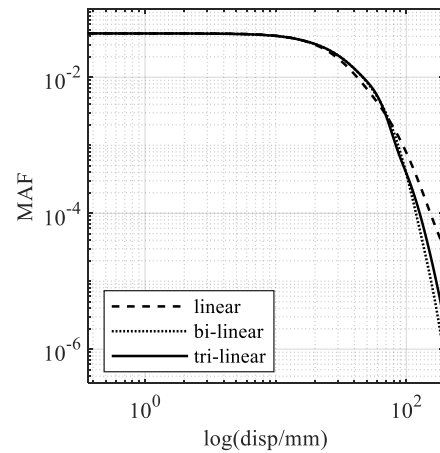


Figure 3: Comparison of the MAF curves based on three different EDP models

Since the EDP formulation has a significant impact on EDP hazard estimation as discussed earlier; therefore, one needs to know the EDP formulation during the adaptive GM selection process in order to check the stop criteria. One solution to this problem is to adopt unsupervised machine learning algorithms to “cluster” data based on a similarity measure. Since EDP usually can be linearly predicted by an IM in logarithm space, a Gaussian mixture model (GMM) for clustering becomes appropriate in this study.

Figure 4 shows the application of GMM clustering to cluster the benchmark data using one, two, and three components, respectively. In order to obtain the best number of clusters, one could use Akaike information criterion (AIC) or Bayesian information criterion (BIC) to determine the best GMM.

The lower value of AIC or BIC indicates a better model. When using the benchmark displacement- S_a data in logarithm space, either AIC or BIC indicates that the one-component model is much worse than the two-components model, and two- and three-components models are about the same. This is consistent with the findings from Figures 1 and 2. Thus, using GMM with AIC or BIC in the adaptive selection process can help determine the appropriate EDP model formulation.

3. ADAPTIVE GM SELECTION

Based on the previous section results, the GM selection for a cloud analysis should address the dependence of EDP-IM relationship on various levels of the conditioning IM. Therefore, it is reasonable to adopt the stratified sampling scheme.

The proposed adaptive GM selection can be illustrated in the flowchart in Figure 5. First the number of bins needs to be determined. For the first iteration, the bins should cover the whole range of the conditioning IM. One could choose the bins with even intervals over the range of IM in logarithm space. Then random choose m records (e.g., $m = 1, 2, \dots$) in each bin, and non-linear time history analysis is conducted based on the chosen records. Within the obtained EDP, the

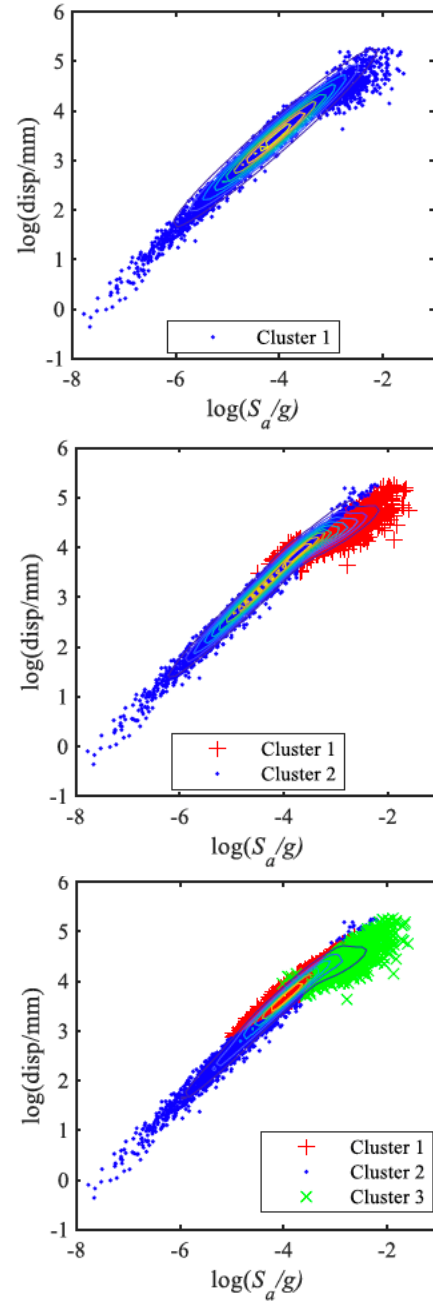


Figure 4: Clustering using Gaussian mixture model using: one component (top), two components (middle), and three components (bottom)

GMM clustering is applied to determine if piecewise regression is needed (that is, to determine if the regression should be linear, bi-linear, tri-linear etc.). Then the model parameters of the EDP model can be determined. This adaptive approach

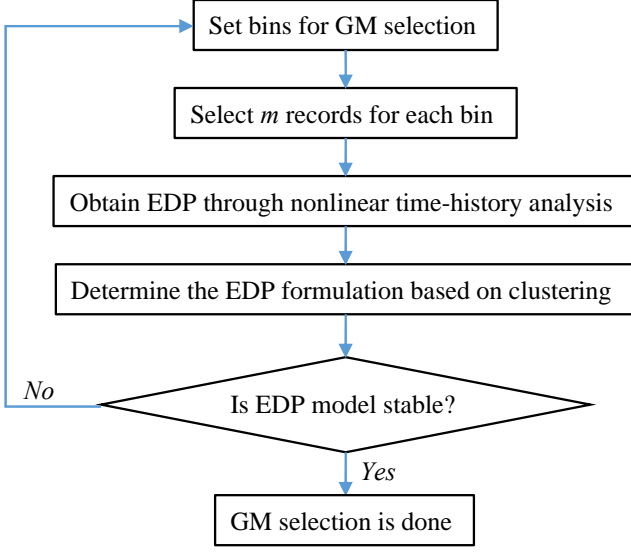


Figure 5: Flowchart of the proposed adaptive GM selection

terminates the iteration when (1) the standard deviation of the model error remains the stable and (2) the marginal error in the estimated demand model coefficients becomes acceptable.

The first stopping criteria can be set up by checking if the model error change from the last iteration to the current iteration is less than a preset tolerance value; the second criteria can be determined using a level of $100(1-\alpha)\%$ confidence interval as follows:

$$\hat{b} \pm e = \hat{b} \pm t_{n-1, \alpha/2} \frac{s}{\sqrt{n}} \quad (4)$$

where \hat{b} is the estimated mean of the model parameter, e is the acceptable marginal error, $t_{n-1, \alpha/2}$ refers to $1-\alpha/2$ quantile of the Student's t distribution with $n-1$ degrees of freedom, s is the estimated standard deviation of the model parameter. Note that when adopting piece-wise regression, the stabilization of each linear segment may not reach at the same time. When this occurs, one could just add GM to the bins that cover the linear segment that has not reached the stabilization in the next iteration.

As an illustration, the adaptive selection is applied by selecting GM records from the benchmark GM set. Then the EDP hazard curve

obtained based on the GM records selected is compared with the benchmark EDP hazard curve to verify the effectiveness of the proposed approach. In particular, 6 bins are used for the adaptive selection and $m = 1$. That is, one from each bin with a total of 6 records is added for each iteration. In addition, it is found before the total number of records (n) reaches 24 (i.e., $n < 24$), the GMM clustering suggests a linear model; and when $n \geq 24$, the GMM clustering suggests a bi-linear model. When $n = 60$, the two stopping criteria are met and the GM selection is terminated.

Figure 6 compares the EDP models based on $n = 24$ and $n = 60$. As expected, the EDP models changes with the number of records selected. In particular, the segment for higher IMs changes significantly when n increases from 24 to 60. This also shows that less records are needed for the first segment assessment compared to the second segment. Figure 7 compares the seismic EDP hazard curves using the selected GMs and EDP hazard curve based on the all benchmark GMs consist of 5000 records. As shown in Figure 7, the hazard curve using 24 records (when the stability of the EDP model is not reached yet) is not able to capture the behavior accurately for large displacements, while using 60 records is able to match the benchmark hazard curve very well overall. These results confirm the effectiveness of the proposed selection process.

4. GM SELECTION FOR TWO IMS

When applying the proposed GM selection, the IM used for the selection is also the IM used for the EDP model development. This is the case when the IM is sufficient for the EDP of interest. However, the GM selection could become challenging, when the IM for the GM selection is not effective or sufficient for the EDP of interest or when there are more than one EDP needed for the analysis and none of single IM could be sufficient for all the EDPs at the same time. Some authors proposed the idea of partitioning EDP-IM space in respect to the considered IMs (e.g., choosing from a grid for two IMs) (Bradley et al. 2015).

In this study, we examine a scenario where one needs to develop the seismic demand models

for two different EDPs with selected GM records, and each EDP has a different effective IM. As one can expect, if the GM selection is based on one IM (say IM_1) and IM_1 is the best proxy for one of the EDP (say EDP_1), IM_1 is not necessary the best proxy for the second EDP (EDP_2). When using a different IM (IM_2) for the seismic demand model development for EDP_2 , the GMs selected based on IM_1 will probably lead to a biased estimation for EDP_2 model. Therefore, the GM selection should also consider some statistical aspects of IM_2 including the correlation of IM_2 with IM_1 .

To study which statistical aspect(s) should be included in the GM selection, two EDPs of the SDOF system described earlier are considered: peak displacement and peak acceleration. In this case, S_a and PGA are the effective IMs for the EDPs, peak displacement and peak acceleration, respectively. First, the proposed adaptive

algorithm is applied to select GMs based on the conditioning IM, S_a , to estimate peak displacement demand model. Then using the selected GMs, the peak acceleration demand model is developed based on PGA. As the GM is randomly selected from the bins, when the proposed GM selection is applied again, a different set of GMs will be generated. For each set of GM, one can examine the statistics of PGA: mean ($\mu_{\ln PGA}$), standard deviation ($\sigma_{\ln PGA}$), and range ($R_{\ln PGA}$) of PGA, and correlation of PGA and S_a ($\rho_{\ln PGA, \ln S_a}$).

Figure 8 shows the effect of PGA statistics on the R-squared (shown in the left plots) and model error standard deviation, σ , (shown in the right plots) of the developed peak acceleration demand model. The dotted lines indicate the linear trend between the PGA statistics and R-squared or σ . As shown in Figure 8, the increases in statistics related to PGA dispersion (i.e., $\sigma_{\ln PGA}$ and $R_{\ln PGA}$) and $\rho_{\ln PGA, \ln S_a}$ increase the demand model R-squared and reduce σ , whereas their effect on sigma is less pronounced. This is particularly true for $\rho_{\ln PGA, \ln S_a}$. This shows the importance to consider the variance of the second IM and the correlation of the first and second IMs in the GM selection.

Furthermore, Figure 9 shows the effect of PGA statistics (i.e., $\mu_{\ln PGA}$, $\sigma_{\ln PGA}$, $R_{\ln PGA}$ and

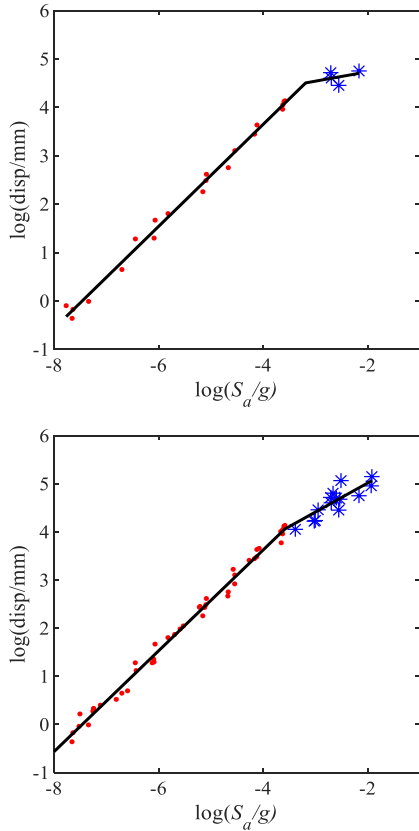


Figure 6: EDP models when $n = 24$ records (top) and when $n = 60$ records (bottom)

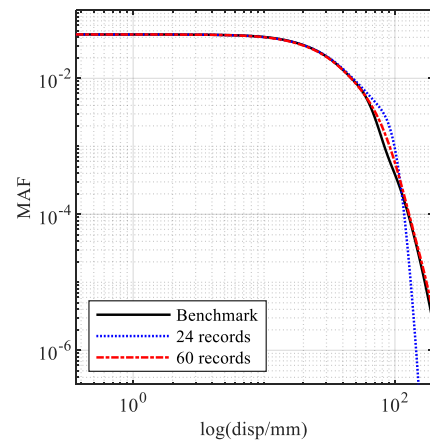


Figure 7: Comparison of the MAF curves based on all benchmark records and the records selected from the proposed GM selection

$\rho_{\ln PGA, \ln Sa}$) on the absolute difference of the demand model coefficients (regression intercept and slope) of the model based on the whole benchmark GMs and the demand model coefficients of the model based on the selected GM, denoted as $|\delta b_0|$ and $|\delta b_1|$, respectively. The smaller value of $|\delta b_0|$ and $|\delta b_1|$ indicates the developed model based on the selected GM is closer to the true behavior. As shown in Figure 9, that the dispersion of PGA (i.e., $\sigma_{\ln PGA}$ and $R_{\ln PGA}$) has impact on the estimation of the demand model slope but the effect is negligible on the model intercept. In addition, as the dispersion of PGA increases, the bias on the estimation of the demand model slope is reduced, suggesting the positive influence of having a widely spread PGA in the GMs. Meanwhile, the

value of $\rho_{\ln PGA, \ln Sa}$ affects both the model intercept and slope significantly.

To illustrate the impact of $\rho_{\ln PGA, \ln Sa}$, Figure 10 compares the demand hazard curves based on the benchmark GMs with the curves based on the two different selected GM sets with two different $\rho_{\ln PGA, \ln Sa}$ values. As shown in Figure 10, the set with higher correlation agrees well with the benchmark. Note that $\rho_{\ln PGA, \ln Sa}$ for the benchmark GMs is 0.74. This further confirms the importance of incorporating $\rho_{\ln PGA, \ln Sa}$ in the GM selection, and one may need to match $\rho_{\ln PGA, \ln Sa}$ with the value in the benchmark GM. To conclude, in order to develop an accurate model for the second EDP, the proposed adaptive algorithm should also incorporate the statistics of the other IM that is effective to the second EDP in the GM selection process.

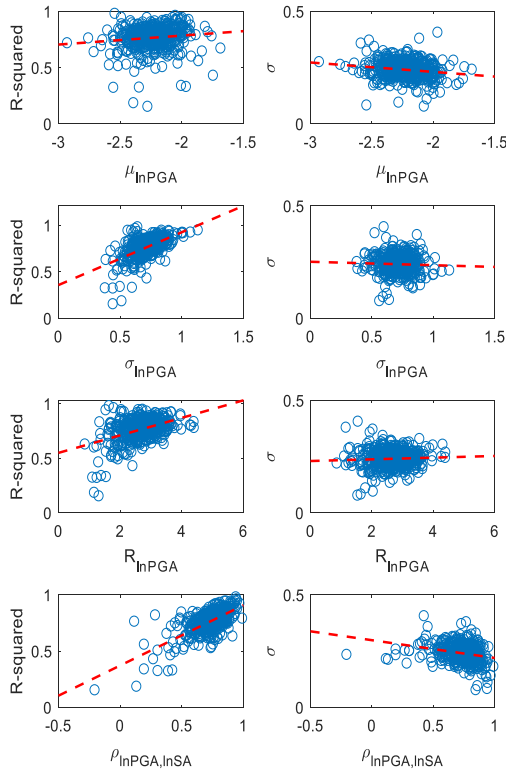


Figure 8: The effect of statistics of PGA on the R-squared and model error of demand model developed for peak acceleration

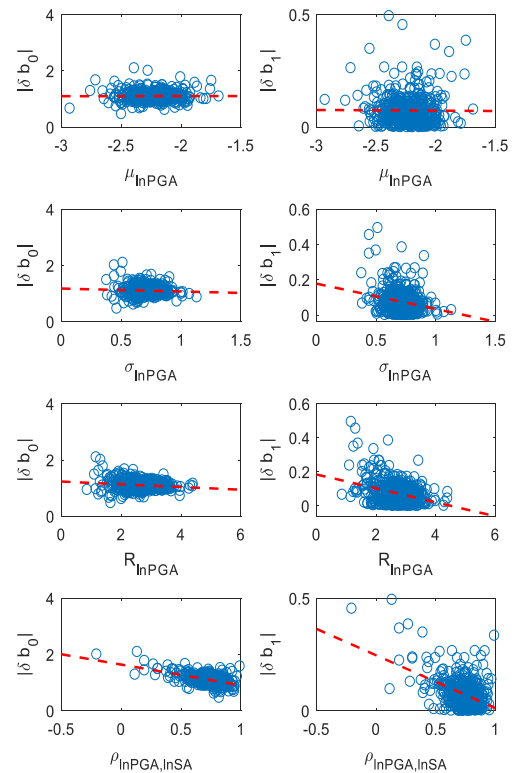


Figure 9: The effect of statistics of PGA on the difference between the demand coefficients of the benchmark model and the demand model based on the selected GMs

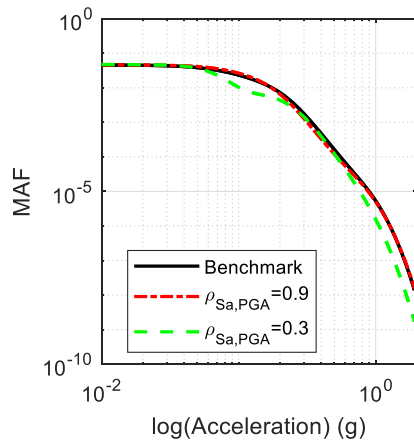


Figure 10: Comparison of the demand hazard curves of two selected sets to the Benchmark

5. SUMMARY AND CONCLUSIONS

In cloud analysis, the structural performance is evaluated by subjecting the numerical model of the structure to a cloud of GMs and regressing the structural responses on the seismic IM. Since the results of cloud analysis are dependent on the input GMs, this study develops an adaptive GM selection procedure to achieve accurate demand model with a small number of time-history analysis.

The developed GM selection uses the stratified sampling scheme to capture the EDP-IM relationship at various levels of IM; it adopts Gaussian mixture model clustering to determine the suitable model formulation for the EDP model. The selection is an iteration process, where a small number of GM records are added at a time and EDP model is checked for each iteration. Once the EDP model is stabilized, the iteration will be terminated. To examine the effectiveness of the GM selection, a nonlinear SDOF system is studied and subjected to a set of site-specific 5000 simulated GMs.

The results show that a piece-wise demand model increases the accuracy of seismic demand hazard curves particular at large EDP levels. It is also found that the EDP hazard curve based on the selected GM records matches very well with the EDP hazard curve based on the benchmark GM records. This validates the proposed selection approach. Lastly, the developed GM selection approach is examined for a scenario where two EDPs models are needed and these two EDPs correspond to two different effective IMs. The

result suggests to apply the proposed adaptive algorithm conditioning on one IM and then choose GMs that have higher dispersion of the second IM and appropriate correlation between the two IMs in each stratum.

REFERENCES

- Bai, J. W., Gardoni, P., & Huete, M. B. D. (2011). Story-specific demand models and seismic fragility estimates for multi-story buildings. *Structural Safety*, 33(1), 96-107.
- Bradley, B. A., Burks, L. S., and Baker, J. W. (2015). "Ground motion selection for simulation-based seismic hazard and structural reliability assessment." *Earthquake Engineering and Structural Dynamics*, 44(13), 2321-2340.
- Jalayer, F., Ebrahimian, H., Miano, A., Manfredi, G., and Sezen, H. (2017). "Analytical fragility assessment using unscaled ground motion records." *Earthquake Engineering and Structural Dynamics*, 46(15), 2639-2663.
- Jalayer, F., De Risi, R., and Manfredi, G. (2015). "Bayesian Cloud Analysis: Efficient structural fragility assessment using linear regression." *Bulletin of Earthquake Engineering*, 13(4), 1183-1203.
- Kwong, N. S. (2015). *Selection and scaling of ground motions for nonlinear response history analysis of buildings in performance-based earthquake engineering*, Doctoral dissertation, UC Berkeley.
- Miano, A., Jalayer, F., Ebrahimian, H., and Prota, A. (2017). "Cloud to IDA: Efficient fragility assessment with limited scaling." *Earthquake Engineering and Structural Dynamics*, 47(5), 1124-1147.
- Watson-Lamprey, J. A. (2007). *Selection and scaling of ground motion time series*. Doctoral dissertation, University of California, Berkeley.
- Zareian, F., Kaviani, P., and Taciroglu, E. (2015). "Multiphase Performance Assessment of Structural Response to Seismic Excitations." *Journal of Structural Engineering*, 141(11), 04015041.